ISSN 1561-5359. Штучний інтелект, 2018, № 3

УДК 004.93+519.25

В.І. Шинкаренко, І.М. Демидович
Дніпровський національний університет залізничного транспорту імені академіка В. Лазаряна, Україна
вул. Лазаряна, 2, м. Дніпро, 49010

ВИЗНАЧЕННЯ ОЗНАК АВТОРСТВА ПРИРОДНОМОВНИХ ТЕКСТІВ

V.I. Shynkarenko, I.M. Demidovich
Dnipro National University of Railway Transport named after academician V. Lazaryan, Ukraine
2, Lazaryan St., Dnipro, 49010

DETERMINATION OF THE ATTRIBUTES OF AUTHORSHIP OF NATURAL TEXTS

Досліджено можливості встановлення авторства природномовних текстів та їх фрагментів методом класифікації за найменшою відстанню у просторі образів. Образи у n-мірному Евклідовому просторі формуються за ознаками вимірювань методами статистичного та рекурентного аналізу, показниками складності тексту. Метод рекурентного аналізу часових рядів адаптовано до аналізу природномовних текстів. Встановлено, що визначені ознаки мають недостатньо високу ефективність при визначенні авторства; у 85% випадків хоча б один з методів дозволяє встановити авторство; модифікований метод рекурентного аналізу має той же рівень ефективності, як статистичний та аналіз складності тексту.

Ключові слова: природномовні тексти, рекурентний аналіз, статистичний аналіз, складність текстів, авторство тексту, класифікація

The possibility of defining the authorship of natural language texts and its fragments was explored by minimum distance classification in space images. In n-dimensional Euclidean space the image forms by measurement signs of statistic and recurrent analysis, complexity indicators. The method of recurrent analysis of time series was adapted to the analysis of natural language texts. Certain signs weren’t efficient enough in authorship determination; in 85% of cases at least one of the methods allows to establish authorship; the modified method of recurrent analysis has the same level of efficiency as statistical and complexity analysis.

Keywords: natural language texts, recurrence analysis, statistic analysis, text complexity, text authorship, classification

Вступ

Науковий інтерес до автоматичної обробки текстів виник приблизно шістдесят років тому. Особливе місце в цій сфері займають проблеми виявлення авторства, плагіату та оцінки якості тексту. На даний час залишається багато невизначеного у цій проблематиці.

Запропонований підхід враховує взаємозв’язок між цими проблемами. Так, замасковані запозичення можуть бути опосередковано виявлені за ознаками авторства та складності текстів.

Постановка проблеми

Виявлення плагіату є однією зі складових у сфері академічної доброчесності. Закон «Про освіту» вимагає перевіряти дисертації, дипломні роботи та наукові публікації на наявність запозичень.

З іншого боку, проблема встановлення авторства текстів виникає у юридичній площині. Питання авторства має велике значення для усіх сфер, де існує поняття права власності на об’єкт, де роль авторства є дуже істотною. Це стосується художніх творів, наукових та навчальних матеріалів та багатьох інших робіт.

Складність питання полягає у тому, що для перевірки текстів на плагіат або виявлення запозичення потрібно мати відповідну базу матеріалів для порівняння. Задача ускладнюється багатомовністю джерел. Частково цю задачу можна вирішити без застосування матеріалів для порівняння.

Аналіз останніх досліджень і публікацій

Частотний аналіз текстів. Проблему статистичної та частотної структури текстів, складання частотних словників мови конкретного автора або окремо взятих текстів на матеріалах різних мов (німецької, англійської, російської і т.д.) досліджували мовознавці [1-4].

Такий аналіз ґрунтується на побудові

частотного словника автора за обраним текстом шляхом обчислення частоти входження кожного зі слововживань [5, 6]. Досвід складання подібних словників наочно демонструє, що словесне наповнення будь-якого, досить довгого тексту, має власну статистичну структуру. Внаслідок чого, можна стверджувати, що у кожного автора є співвідношення часто і рідко вживаних лексем. Саме це співвідношення читач і сприймає як багатий чи бідний словник автора [7, 8].

У подальшому, після проведення частотного аналізу, виділяються визначальні ознаки для кожного з текстів. Однією з таких характеристик є авторський інваріант [9]. Це – числовий параметр, який дає можливість розрізняти твір за авторським стилем. Дуже часто, як показали попередні дослідження для прози, на цей показник істотно впливає частота вживання службових слів (прийменники, сполучники або частки).

Частотним характеристикам текстів присвячено багато робіт, де були розглянуті подібності між авторами XIX-XX століть [10]. Також були проаналізовані подібні словники для різних слов'янських мов, таких як чеська, польська, сербська, болгарська та російська [11].

Аналіз на основі N-грам. Одним з широко використовуваних методів аналізу тексту є метод N-грам [12]. Він є часто вживаним у виявленні плагіату [13]. Цей метод став застосовуватися порівняно недавно.

N-грамом в алфавіті називають довільний ланцюжок довжиною N. Як ланки такого ланцюжка можна використовувати як символи, так і окремі слова. Метод полягає у підрахунку і порівнянні профілів частоти N-грамів для різних текстів.

У багатьох задачах необхідно визначити, так званий, стиль тексту. Під стилем тексту розуміється сформована система мовних засобів, використовуваних у різних сферах людського спілкування. У лінгвістиці його прийнято називати функціональним стилем мови [14]. Стиль тексту багато в чому визначається частотою і порядком вживання у ньому різних частин мови [14], що задовольняє умовами застосування методу N-грам.

Аналіз на основі N-грам дає можливість виявити характерні сполучення слів та їх складність для конкретного твору або автора. На основі цих даних можна визначити характерний стиль мовлення автора. Дане твердження справедливе як для звичайних, так і для спеціалізованих текстів [15].

Показники складності сприйняття тексту. Лексику прийнято вважати найкращим показником легкості сприйняття тексту. Середня довжина слів (у буквах або символах) і речень є статистичними факторами, які часто використовують для оцінки складності тексту. Ці параметри легко піддаються кількісному вираженню і придатні для автоматичної оцінки.

Проблему визначення складності тексту для розуміння читачем допомагають вирішити цілий ряд показників. Наприклад, індекси туманності Ганнінга, Колемана-Лиау та оцінка читабельності Рейгора [16]. Вони будуються на основі підрахунку кількості речень, слів, складів, букв у тексті, також середньої кількості слів, складів, букв у реченнях та складів і букв у словах.

Усі перелічені вище показники розраховувались для текстів англійської мови вузького призначення та для певної аудиторії читачів [16]. Тому вони не зовсім відповідають меті дослідження, однак початкові кількісні показники мають певну інформативність.

Ступінь складності текстів може давати відповідну характеристику автору.

Рекурентний аналіз. Рекурентний аналіз використовується для дослідження часових рядів. Він був модифікований для аналізу текстів.

За основу був узятий аналіз рекурентних діаграм (recurrence quantification analysis, RQA), у якому для аналізу використовують щільність рекурентних точок [17].

Модифікований метод полягає у наступному:

- розраховується частота букв у тексті;
- отримується часовий ряд, замінюючи кожний символ обраного тексту на його частоту. Умовний час – перехід від одного символу до іншого;
- визначається фазовий простір [18], як візуалізація переходів від стану до стану (від символа до символа);
- розраховується рекурентна діаграма на основі фазового простору через відображення повторюваних станів у різні моменти часу;
- обчислюються та інтерпретуються загальновживані показники рекурентного аналізу щодо аналізу тексту.

Показник рекурентності (recurrence rate, *RR*) визначає щільність рекурентних точок на досліджуваній діаграмі. Це значення приблизно відображає загальну кількість повторень кожного зі статистично близьких символів

$$RR = \frac{1}{N^2}\sum_{i,j=1}^{N} R_{i,j}^{m,\varepsilon}, \qquad (1)$$

де *N* – кількість розглянутих станів, $R_{i,j}$ – *i,j*-та точка рекурентної діаграми, ε – радіус околиці точки в момент часу *i*, *m* – розмірність фазового простору.

Показник детермінізму (determinism, *DET*) розглядає частотний розподіл довжин *l* діагональних ліній у діаграмі $P^{\varepsilon}(l)$, де *N* – абсолютна кількість таких ліній. Значення *DET* визначає частоту повторень усіх сполучень статистично близьких символів будь-якої довжини:

$$DET = \frac{\sum_{l=l_{\min}}^{N} lP^{\varepsilon}(l)}{\sum_{i,j}^{N} \mathrm{R}_{\mathrm{i,j}}^{m,\varepsilon}}. \qquad (2)$$

Середня довжина діагональних ліній *L* визначає середню довжину повторюваних статистично близьких символів.

$$L = \frac{\sum_{l=l_{\min}}^{N} lP^{\varepsilon}(l)}{\sum_{l=l_{\min}}^{N} P^{\varepsilon}(l)}. \qquad (3)$$

Показник дивергенції (divergence, *DIV*) є величиною, зворотною максимальній довжині діагональних структур.

$$DIV = \frac{1}{\max(\{l_i; i = 1...N_l\})}. \qquad (4)$$

Ентропія (entropy, *ENTR*) є показником частотного розподілу діагональних ліній, для текстів – частотного розподілу повторюваних поєднань статистично близьких символів.

$$ENTR = -\sum_{l=l_{\min}}^{N} p(l)\ln(p), \qquad (5)$$

де

$$p(l) = \frac{P^{\varepsilon}(l)}{\sum_{l=l_{\min}}^{N} P^{\varepsilon}(l)}. \qquad (6)$$

Показник завмирання (laminarity, *LAM*) демонструє частотний розподіл довжин v-горизонтальних ліній у діаграмі $P^{\varepsilon}(v)$, де *N* – абсолютна кількість таких ліній. Показник *LAM* приблизно визначає повторення статистично близьких символів.

$$LAM = \frac{\sum_{v=v_{\min}}^{N} vP^{\varepsilon}(v)}{\sum_{i,j}^{N} \mathrm{R}_{\mathrm{i,j}}^{m,\varepsilon}}. \qquad (7)$$

Показник затримки (trapping time, *TT*) відображає середню довжину горизонтальних структур. Показник *TT* визначає середню довжину поєднань статистично близьких символів.

$$TT = \frac{\sum_{v=v_{\min}}^{N} vP^{\varepsilon}(v)}{\sum_{v=v_{\min}}^{N} P^{\varepsilon}(v)}. \qquad (8)$$

Показники (1)..(8) відображають структуру рекурентної діаграми.

Мета дослідження

Задача даної роботи полягає у визначенні ефективності методів статистичного та рекурентного аналізу, показників складності тексту щодо встановлення авторства текстів.

Експериментальні дослідження ефективності ознак авторства

Підготовка експерименту. Для проведення експерименту була обрана художня література через її яскраво виражену індивідуальність та достовірність інформації про авторство.

Для коректного проведення експерименту була сформована навчальна вибірка з 20 творів 11 авторів та контрольна вибірка з 33 текстових файлів: по три тексти кожного автора з навчальної вибірки.

Визначення параметрів рекурентного аналізу. Спочатку наведемо реалізацію модифікованого методу рекурентного аналізу [18] на прикладі «Заповіту» Т. Шевченка (рис. 1-4).

Обчислення частоти входження кожного символу українського алфавіту наведені на рис. 1 у вигляді стовпчикової діаграми.

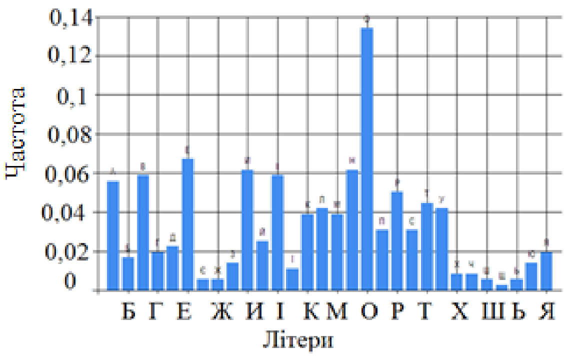


Рис. 1. Діаграма з частотою символів

На рис. 2 представлено часовий ряд, сформований на основі обраного тексту з відповідними (як на рис. 1) частотами.

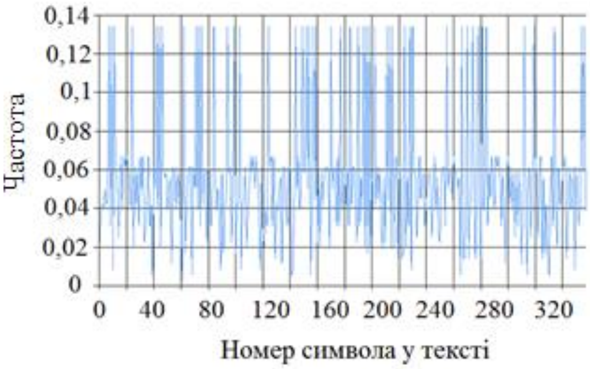


Рис. 2. Часовий ряд тексту «Заповіт»

За отриманими частотами відповідно до всього тексту «Заповіту» за канонами рекурентного аналізу [18] визначено фазовий простір (рис. 3) розмірністю – 2.

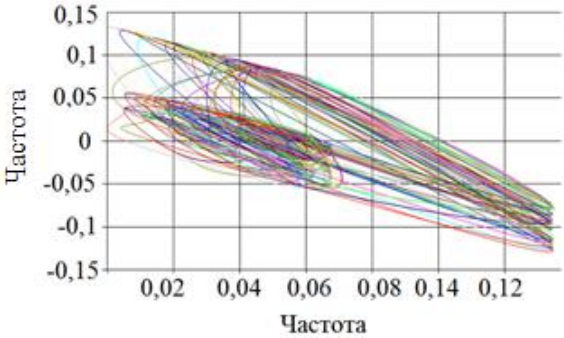


Рис. 3. Фазовий простір тексту

Побудована рекурентна діаграма має відображати особливості авторського тексту. Діаграма згідно з «Заповітом» наведена на рис. 4. Значення радіусу околиці точок у фазовому просторі $\varepsilon = 0,5$.

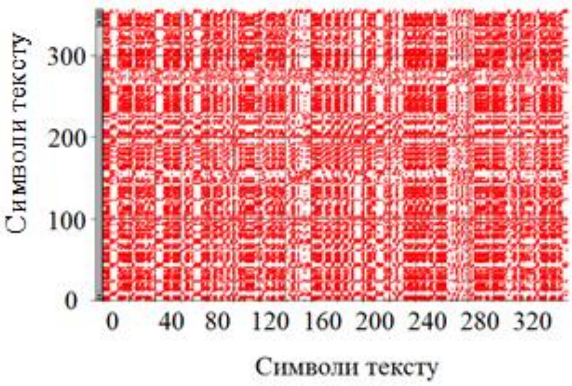


Рис. 4. Рекурентна діаграма тексту

Для спрощення аналізу діаграми обраховуються показники рекурентності (1)..(8). Для «Заповіту» отримані наступні значення показників (табл. 1).

Таблиця 1. Показники рекурентного аналізу «Заповіту»

Назва показників	*Значення*
Міра рекурентності RR	0,021
Міра детермінізму DET	0,002
Дивергенція DIV	0,125
Середня довжина діагоналей L	2,38
Міра ентропії ENTR	0,769
Міра завмирання LAM	0,00018
Міра затримки TT	2

Виконання експерименту. Виконані експериментальні дослідження тексту за частотою літер, довжиною слів та рекурентним аналізом.

Авторство тексту визначається за найменшою відстанню до еталону [19], за який приймається середнє значення за творами автора з навчальної вибірки.

Вважаємо, що образ X_{ij} належить до класу ω_k, якщо найближчий до X_{ij} образ навчальної вибірки належить ω_i (X_{ij} – вектори у Евклідовому просторі, де *i* – показник, за яким визначається авторство, *j* – номер твору в навчальній або контрольній вибірці, x_{ijk} – *k*-ий елемент вектору X_{ij}).

Класифікування виконується окремо за частотою літер (X_{1j}), довжиною слів (X_{2j}), показниками рекурентного аналізу (X_{3j}), та усіма показниками разом (X_{4j}). Елементи векторів x_{1jk} – *k*-ий показник рекурентного аналізу (табл. 1); x_{2jk} – частота *k*-ї літери у тексті; x_{3jk} – кількість слів довжиною *k*-літер.

Згідно з текстом «Заповіту» отримані значення векторів X_{1j}, X_{2j}, X_{3j}, X_{4j}.

$X_{1,1} = [0.02 \quad 0 \quad 0.13 \quad 2.38 \quad 0.77 \quad 0 \quad 2]$;

$X_{2,1} = [0.06 \quad 0.02 \quad 0.06 \quad 0.02 \quad ... \quad 0.02]$;

$X_{3,1} = [14 \quad 11 \quad 2 \quad 18 \quad 11 \quad 11 \quad 7 \quad 4 \quad 4 \quad 1]$;

$X_{4,1} =$ [0.02 0 0.13 2.38 0.77 0 2 0.06 0.02 … 0.02 14 11 2 18 11 11 7 4 4 1].

Для коректності порівняння вектори були унормовані наступним чином:

$$x^*_{ijk} = \frac{x_{ijk} - \min_j(x_{ijk})}{\max_j(x_{ijk}) - \min_j(x_{ijk})}.$$

У результаті обробки контрольної вибірки були отримані результати, наведені у табл. 2, де сірим виділені ті результати, що виявили автора твору, або були близькі до нього.

Авторство творів у таблиці подано наступним чином: 1 – О. Довженко, 2 – І. Багряний, 3 – І. Франко, 4 – М. Коцюбинський, 5 – Л. Українка, 6 – М. Хвильовий, 7 – О. Вишня, 8 – П. Мирний, 9 – В. Підмогильний, 10 – С. Жадан, 11 – Т. Шевченко.

Таблиця 2. Визначення авторства текстів з використанням аналізу за одним символом

Автор	*ЧЛ*	*ЛС*	*РА*	*Загальне*
2	2	9,2,6/5	6,2,7/15	2
2	8,1,4/10	11,10,5/75	1,2,7/31	10,11,7/27
2	6,9,4/16	9,8,6/36	1,6,9/24	9,8,6/29
7	6,8,2/14	6,9,3/21	8,4,3/9	6,8,9/17
7	2,6,4/13	2,1,7/20	6,7,9/2	2,1,7/16
7	8,7,4/4	7	4,3,9/23	7
1	8,3,9/14	11,7,10/42	8,4,11/39	7,4,11/13
1	1	6,9,8/50	6,2,1/15	6,9,8/32
1	3,7,8/17	8,3,6/17	1	3,8,9/16
10	2,6,10/9	9,6,8/54	1,2,9/33	9,6,2/49
10	10	10	10	10
10	1,9,2/19	10	10	1,10,4/3
4	4	6,9,8/20	8,6,4/36	6,9,8/19
4	1,4,2/5	6,9,3/30	6,9,1/30	6,8,9/20
4	9,1,4/1	8,9,6/22	7,3,6/50	9,8,4/4
5	6,1,4/17	1,4,8/70	8,4,11/19	4,8,1/40
5	4,7,5/7	4,3,7/74	6,7,2/60	4,3,7/48
5	5	11,4,10/65	10,7,1/47	11,5,10/1
8	4,8,3/15	10,11,7/61	8	10,11,7/27
8	8	8	10,1,9/52	8
8	8	8	2,1,9/29	8
9	6,9,3/1	5,10,11/21	10,1,2/32	10,5,11/55
9	1,4,9/8	1,2,7/40	10,1,9/18	1,2,4/30
9	9	1,2,4/39	4,9,1/22	1,2,4/26
3	2,4,9/10	2,1,6/33	1,7,9/15	2,1,9/43
3	5,4,1/8	11,5,10/80	10,1,2/57	5,10,11/40
3	1,9,4/24	8,9,6/25	2,1,9/22	9,6,5/10
6	4,6,1/4	4,3,1/48	6	4,1,3/23
6	6	7,1,4/56	6	7,1,4/31
6	6	9,6,8/8	7,1,2/14	9,6,2/7
11	10,7,1/18	9,6,8/51	6,3,9/78	6,9,8/42
11	11	8,3,4/73	2,7,9/10	8,3,4/42
11	11	11	11	11

Інші стовпчики у табл. 2: ЧЛ (частота літер – за вектором X_{2j}); ЛС (кількість літер у слові – за X_{3j}); РА (рекурентний аналіз – за X_{1j}); загальне – результати порівняння за об'єднаним вектором X_{4j}.

У комірках таблиці – інформація що-

до визначення найближчих трьох авторів для обраного твору. Якщо перший результат є точним, то наступні не наводяться. Четверте значення визначає близькість першого отриманого результату до реального авторства наступним чином:

$$p = \frac{|l_2 - l_1|}{\max(l_1, l_2)},$$

де l_1 – відстань між векторами твору та найближчим еталоном, l_2 – відстань між векторами твору та еталоном творів реального автора.

Також було виконано визначення автора тексту з використанням N-грамів. Цей метод заснований на розбитті усього тексту на пари сусідніх символів та визначенні їх частоти, з якою вони зустрічаються у творі. При цьому до пари входять символи з нахлестом, тобто спочатку обираються перший та другий символи, потім другий та третій і т.д. Якщо у слові залишається лише один символ, то в пару до нього йде перший символ наступного слова.

Були проведені експерименти для 2- ... 7-грамів із заміною поетичних творів на прозові.

Авторство творів у таблицях 3, 4 пронумеровано наступним чином: 1 – І. Багряний, 2 – О. Вишня, 3 – М. Вовчок, 4 – О. Довженко, 5 — М. Коцюбинський, 6 – Г. Квітка-Основ'яненко, 7 – П. Мирний, 8 – В. Нестайко, 9 – В. Підмогильний, 10 – І. Франко, 11 – М. Хвильовий.

Найкращий результат був отриманий при застосуванні 4-грамів (табл. 3).

Аналіз даних у табл. 3 щодо встановлення авторства за допомогою 4-грамів виявив суттєве покращення аналізу з використанням частоти символів, але зменшення ефективності використання рекурентного аналізу.

Також було виконане порівняння за частотою слів з урахуванням їх закінчень.

Другий стовпчик табл. 4 – ЧС (результати порівняння за вектором X_{1j} з даними частоти слів у тексті).

Для виявлення авторства розрахована частота усіх слів у тексті з подальшим формуванням часового ряду, фазового простору та рекурентної діаграми за отриманими даними (табл. 4).

Таблиця 3. Визначення авторства текстів за 4-грамами

Автор	*ЧЛ*	*ЛС*	*РА*	*Загальне*
1	1	8	2	1
1	1	7	5	1
1	1	9	9	1
2	2	4	6	2
2	2	8	9	2
2	2	2	5	2
3	3	3	2	3
3	3	7	6	3
3	3	3	3	3
4	4	10	5	4
4	4	7	8	4
4	4	6	8	4
5	5	5	3	5
5	5	5	11	5
5	5	7	3	5
6	6	6	7	6
6	6	6	3	6
6	6	6	9	6
7	4	9	2	4
7	7	9	3	7
7	7	7	7	7
8	8	9	11	8
8	8	5	4	8
8	8	8	11	8
9	2	2	5	2
9	9	1	5	9
9	9	1	5	9
10	1	1	2	1
10	5	10	5	5
10	10	5	9	10
11	11	7	2	11
11	11	4	2	11
11	9	1	3	9

Дані табл. 4 дозволяють стверджувати, що встановлення авторства твору з ви-

користанням частоти слів дещо гірше за ефективність аналізу по 4-грамам.

Таблиця 4. Визначення авторства текстів за словами

Автор	*ЧЛ*	*ЛС*	*РА*	*Загальне*
1	1	8	1	1
1	2	7	9	2
1	1	9	10	1
2	2	4	11	2
2	2	8	6	2
2	2	2	6	2
3	3	3	8	3
3	3	7	2	3
3	3	3	3	3
4	4	10	6	4
4	4	7	1	4
4	2	6	7	2
5	5	5	1	5
5	5	5	10	5
5	5	7	8	5
6	6	6	7	6
6	6	6	11	6
6	6	6	2	6
7	2	9	3	2
7	7	9	2	7
7	7	7	7	7
8	8	9	11	8
8	2	5	2	2
8	8	8	2	8
9	2	5	1	2
9	9	1	10	9
9	9	1	1	9
10	2	1	1	2
10	2	10	7	2
10	10	5	10	10
11	11	7	2	11
11	2	4	6	2
11	11	1	3	11

Висновки

При визначенні авторства текстів контрольної вибірки при першому проведенні експерименту безпомилково визначилися лише автори 2 текстів. Кращий результат визначення авторства дав метод з використанням частоти букв – 12 збігів по автору. Решта методів визначили автора всього у 6 випадках та у 7 за даними рекурентного аналізу.

Відсоток близькості знаходиться у широкому діапазоні від 1% до 80%. Окремо за методами: за даними про частоту літер – 24%, для даних щодо кількості літер у словах – 80%, для рекурентного аналізу тексту – 78% та за результатами порівняння з використанням усіх отриманих даних – 55%.

Також у 22 випадках аналізу тексту автор визначався другим або третім за відстанню. Найкращий показник також за даними щодо частоти літер у тексті, а наступний – за показниками рекурентного аналізу.

Найкращі результати були отримані при визначенні авторства творів за допомогою 4-грамів та по словах – 85 % та 76 % відповідно за загальним вектором.

Покращення результатів слід очікувати при розширені методів класифікації за обраними показниками, враховуючи словосполучення та частини слів. Не досліджено можливості попередньої обробки часових рядів та кодування отриманих даних.

Література

1. Wimmer, G., Altmann, G., Hřebíček, L, Ondrejovič, S., Wimmerová, S. (2003) *Úvod do analýzy textov. Bratislava*, – 344 p.
2. Popesku, I.I., Altmann, G. (2006) Some aspects of word frequencies. *Glottometrics*. №13, – P. 23-46.
3. Köhler, R., Altmann, G. (2005) Aims and Methods of Quantitative Linguistics. *Problems of Quantitative Linguistics*. Chernivci, – P. 12-42.
4. Перебийніс, В.С. (2002) *Статистичні методи для лінгвістів*: Навчальний посібник. Вінниця, – 168 с.
5. Alekseev, P.M. (2005) *Frequency dictionaries.* Quantitative Linguistik : ein internationales Handbuch = Quantitative linguistics : an international handbook/ edited by Reinhard Kohler, Gabriel Altmann, Rajmund G. Piotrowski. Berlin – New York. – P. 312-324.
6. Popescu, I. (2009) *Word frequency studies.* Berlin– New York, – 276 p.
7. Сухорольська, С.М., Федоренко, О.І. (2009) *Методи лінгвістичних досліджень:* Навч. посібник. Львів, – 348 с.
8. Чатуев, М.Б., Чеповский, А.М. (2011) *Частотные методы в компьютерной лингвистике.* – М.: МГУП. – 88 с.
9. Фоменко, В.П., Фоменко, Т.Г. (1996). Авторский инвариант русских литературных текстов. *Новая хронология Греции: Античность в средневековье.* Т. 2. М.: Изд-во МГУ, – С. 768-820.

10. Баевский, В.С. (2001) *Лингвистические, математические, семиотические и компьютерные модели в истории и теории литературы.* М., – 312 с.
11. Бук, С. (2011) *Слов'янський досвід укладання частотних словників мови письменника*. *Проблеми слов'янознавства.* Львів, – С. 217-224.
12. Бузикашвили, Н.Е., Самойлов, Д.В., Крылова, Г.А. (2000) N-граммы в лингвистике. Сборник: *Методы и средства работы с документами*. М.: Диториал УРРС, – 376 с.
13. Тарануха, В.Ю. (2014) Использование комбинированных критериев для автоматизированного определения заимствований. «*Инновации в науке»: сборник статей по материалам XXXII международной научно-практической конференции*. Новосибирск: Изд. «СибАК». – С. 15-18.
14. Кожина, М.Н., Дускаева, Л.Р., Салимовский, В.А. (2008) *Стилистика русского языка*. М.: Флинта: Наука. – 464 с.
15. William, B., Cavnar, John M. (1994) Trenkle *N-Gram-Based Text Categorization.* Michigan, – P. 161–175.
16. Рогушина, Ю.В. (2007) Использование критериев оценки удобочитаемости текста для поиска информации, соответствующей реальным потребностям пользователя. *Проблеми програмування*. Київ, – С. 76-88.
17. Zbilut, J.P., Webber, Jr.C.L. (1992) Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*.– V.171. № 3-4. – P. 199–203.
18. Ту, Дж., Гонсалес, Р. (1978) *Принципы распознавания образов.* М., – 411 с.
19. Киселев, В.Б. (2006) Рекуррентный анализ – теория и практика. *Научно-технический вестник информационных технологий, механики и оптики*. №29, – СПб. – С. 118-127.

References

1. Wimmer, G., Altmann, G., Hřebíček, L, Ondrejovič, S., Wimmerová, S. (2003) Úvod do analýzy textov. Bratislava, – 344 p.
2. Popesku, I.I., Altmann, G. (2006) Some aspects of word frequencies. *Glottometrics*. №13, – P. 23-46.
3. Köhler, R., Altmann, G. (2005) Aims and Methods of Quantitative Linguistics. *Problems of Quantitative Linguistics*. Chernivci, – P. 12-42.
4. Perebyjnis, V.S. (2002) *Statystychni metody dlja lingvistiv*: Navchal'nyj posibnyk. Vinnycja, – 168 s.
5. Alekseev, P.M. (2005) *Frequency dictionaries.* Quantitative Linguistik : ein internationales Handbuch = Quantitative linguistics : an international handbook/ edited by Reinhard Kohler, Gabriel Altmann, Rajmund G. Piotrowski. Berlin – New York. – P. 312-324.
6. Popescu, I. (2009) *Word frequency studies.* Berlin–New York, – 276 p.
7. Suhorol's'ka, S.M., Fedorenko, O.I. (2009) *Metody lingvistychnyh doslidzhen*': Navch. posibnyk. L'viv, – 348 s.
8. Chatuev, M.B., Chepovskii, A.M. (2011) *Chastotnye metody v komp'yuternoi lingvistike*. – M.: MGUP. – 88 s.
9. Fomenko, V.P., Fomenko, T.G. (1996) Avtorskii invariant russkikh literaturnykh tekstov. *Novaya khronologiya Gretsii: Antichnost' v srednevekov'e*. T. 2. M.: Izd-vo MGU, – S.768-820.
10. Baevskii, V.S. (2001) *Lingvisticheskie, matematicheskie, simeoticheskie i komp'yuternye modeli v istorii i teorii literatury*. M., – 312 s.
11. Buk, S. (2011) *Slov'jans'kyj dosvid ukladannja chastotnyh slovnykiv movy pys'mennyka*. *Problemy slov'janoznavstva*. L'viv, – S. 217-224.
12. Buzikashvili, N.E., Samoylov, D.V., Kryilova, G.A. (2000) N-grammyi v lingvistike. Sbornik*: Metodyi i sredstva rabotyi s dokumentami*. M.: Ditorial URRS, – 376 s.
13. Taranuha, V.Yu. (2014) Ispolzovanie kombinirovannyih kriteriev dlya avtomatizirovannogo opredeleniya zaimstvovaniy. *«Innovatsii v nauke»: sbornik statey po materialam XXXII mezhdunarodnoy nauchno-prakticheskoy konferentsii*. Novosibirsk: Izd. «SibAK». – S. 15-18.
14. Kozhina, M.N., Duskaeva, L.R., Salimovskiy, V.A. (2008) *Stilistika russkogo yazyika*. M.: Flinta: Nauka. – 464 s.
15. William, B. Cavnar, John M. (1994) Trenkle *N-Gram-Based Text Categorization.* Michigan, – P. 161–175.
16. Rogushina, Yu.V. (2007) Ispolzovanie kriteriev otsenki udobochitaemosti teksta dlya poiska informatsii, sootvetstvuyuschey realnyim potrebnostyam polzovatelya. *Problemi programyuvannya*. Kyiv, – S. 76-88.
17. Zbilut, J.P., Webber, Jr.C.L. (1992) Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*.– V.171. № 3-4. – P. 199–203.
18. Tu, Dzh., Gonsales, R. (1978) *Printsipyi raspoznavaniya obrazov*. M., – 411 s.
19. Kiselev, V.B. (2006) Rekurrentnyiy analiz – teoriya i praktika. *Nauchno-tehnicheskiy vestnik informatsionnyih tehnologiy, mehaniki i optiki*. №29, – SPb. – S. 118-127.

RESUME

V.I. Shynkarenko, I.M. Demidovich

Determination of the attributes of authorship of natural texts

The research has been done in the field of intellectual processing of natural language texts and their fragments.

The purpose of this work is to define the effectiveness of statistical and recurrent analysis methods, and text complexity indicators to determine the authorship of texts and their fragments, as well as to reveal the plagiarism suspicions.

The parameters for solving these problems were frequency of symbols in texts, indicators of recurrent analysis and text complexity.

The method of recurrent analysis of time series has been adapted for natural language analysis.

Four groups were formed to determine the efficiency of each parameter. The first group has symbols frequency data, the second – words length data, the third – recurrent analysis data and the fourth group has aggregated data for all three previous groups.

The training and control samples have been formed from 11 Ukrainian fiction authors. This type of literature was chosen because of its strongly marked individuality and reliable information about its authorship. For each of the authors the standard has been calculated – the average values for all of previous parameters.

The received images of texts from control sample were classified by the method of minimum distance to the standard for all previous parameters in the Euclidian space of images.

Texts were processed by the following ways: character by character, 2-...7-grams and words with its suffix.

It was established that certain signs weren't efficient enough in authorship determining. Only in 85% of cases at least one of the methods allows to establish the author.

The modified method of recurrent analysis has the same level of efficiency as statistical and complexity analyzes using the text symbols frequency, and slightly lower using N-grams and words analyzes.

The using 4-grams have been the most effective method in authorship determination.

Improvement of results should be expected with expanded classification methods based on selected parameters, including words combinations and parts of words. The possibilities of advanced processing of time series and coding of the received data are not investigated.

Надійшла до редакції 18.10.2018

